LIBRARY
OF THE
MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

Multivariate Measures of Profile Similarity for the Objective Stratification of Econometric Data

251-67

MASS. INST. TECH.
APR 25 1967
LIBRARY

Donald E. Farrar*

*Sloan School of Management, Massachusetts Institute of Technology. This paper condenses part of a larger study by the author in collaboration with John R. Meyer and Robert R. Glauber. The author would like to thank his collaborators, as well as Gordon M. Kaufman of the Sloan School of Management and Kenneth J. Jones of Brandeis University for their considerable interest and assistance in the research on which the paper is based.

This study has been financed by grants for research in Finance from the Ford Foundation to the author directly, and to the Sloan School of Management, Massachusetts Institute of Technology.

RECEIVED
APR 26 1967
M. I. T. LIBRARIES

Multivariate Measures of Profile Similarity for the Objective Stratification of Econometric Data

by

Donald E. Farrar

Motivation

A common problem in the econometric analysis of cross sectional data is to control by stratification for interfirm differences that cannot be handled by adding explanatory variables or equations directly to a structural model.[1]

Consider, for example, two firms that experience an identical (say 20%) increase in sales. The first may be an aggressive member of a rapidly and steadily growing industry, the second a conservative, marginal producer in a highly cyclical market. The first firm's management is likely to interpret an increase in sales as further confirmation of the company's underlying market strength, while the latter may attach little permanence to an isolated upswing. The first, accordingly, may respond to the stimuli by increasing expenditures on plant and equipment to meet anticipated capacity needs, while the latter may actually reduce such outlays to fund the increases in working capital required to support a higher level of sales.

Econometric studies of the relationship between sales, working capital and investment based on cross sections of data containing firms such as

---

[1] See J. R. Meyer, "On the Measurement of Business Motivation," Review of Economics and Statistics, 49, 3, 1967; R. Eisner and R. H. Strotz, "Determinants of Business Investment," Impacts of Monetary Policy, Commission on Money and Credit, Prentice-Hall, 1963; and R. Eisner, "A Permanent Income Theory for Investment," forthcoming in the American Economic Review.

these (in approximately equal numbers) are almost certain to indicate that no relationship whatever exists between investment and either sales or working capital. In fact, of course, the problem is not that no relationship exists between these variables, but rather that two quite different relationships exist, each of which is perfectly understandable in its own context, and quite consistent with the character and the environment of the firms in which the behavior is observed. Unfortunately, when two distinctly different types of behavior appear in the same cross section of data neither is likely to be detected, for the data generated by one may largely wash-out the statistical implications of that generated by the other.[2] The net effect of mixing such firms in the same pool of data, therefore, is to obscure two perfectly meaningful theories rather than to uncover either; leaving only a third, bad theory (in many cases a null hypothesis) to explain observed or to forecast future behavior.

Requirements for Effective Stratification

For econometric purposes, therefore, our objective must be to stratify a sample of observations into subsets that are both useful for the purpose at hand (explaining and forecasting economic behavior) and meaningful in terms of the basic "characters" of the groups obtained. More formally, let us define:

---

[2] As Theil demonstrates, a sample containing observations generated by two or more underlying populations will yield regression coefficients that are, simply, estimates of a weighted average of the different underlying structural parameters; H. Theil, Linear Aggregation of Economic Relations, North Holland, Amsterdam, 1954; and _____, Economic Forecasts and Policy, 2nd Revised Edition, North Holland, Amsterdam, 1961.

$\underline{\underline{Z}}$ as an Nxp matrix of N observations on each of p "profile variables" that span a space within which each observation's underlying "character" is defined,

$\underline{\underline{Y}}$ as an Nxr set of N observations on r endogenous, or jointly dependent variables,

$\underline{\underline{X}}$ as an Nxn array of N observations on n exogenous variables, and

$\underline{\underline{U}}$ as an Nxr set of true, unobserved stochastic error terms corresponding to $\underline{\underline{Y}}$.

Should a sample of observations contain more than one "behavioral type," however, it will not be possible to assume that a single set of nxr structural parameters $\underline{\underline{\beta}}$ exists that adequately represents a (presumably linear) structural relationship of the type generally postulated between $\underline{\underline{Y}}$ and $\underline{\underline{X}}$,

$$\underline{\underline{Y}} = \underline{\underline{X}}\,\underline{\underline{\beta}} + \underline{\underline{U}}. \tag{1}$$

Instead, it will be necessary to stratify existing data into subsets such that, for each set s=1,2,...,S, both of two requirements are satisfied:

1. Behavioral Homogeneity: the relationship between $\underline{\underline{Y}}^{(s)}$ and $\underline{\underline{X}}^{(s)}$,

$$\underline{\underline{Y}}^{(s)} = \underline{\underline{X}}^{(s)}\underline{\underline{\beta}}^{(s)} + \underline{\underline{U}}^{(s)} \tag{2}$$

satisfies the usual statistical properties of a general linear model[3] within set s; although not necessarily across sets, and

2. Profile Homogeneity: the set of observations is partitioned into contiguous subsets $\underline{\underline{Z}}^{(1)},\underline{\underline{Z}}^{(2)},\ldots,\underline{\underline{Z}}^{(S)}$, such that on the basis of

---

[3] F. Graybill, An Introduction to Linear Statistical Models, Volume 1, McGraw-Hill, New York, 1961.

an appropriately defined metric in $\underline{Z}$, each observation is closer to its own centroid than to any other, and group centroids are sufficiently distinct to permit meaningful interpretation.

No explanation beyond our initial illustration and Theil's demonstration of the weighted average character of estimated regression coefficients over behaviorally mixed samples is needed to justify the first, behavioral homogeneity requirement. As for the second requirement, consider the possibility of employing behavior directly to stratify a set of firms according to alternative patterns of investment behavior. Firms whose investment varies directly with sales, perhaps, could be stratified into an aggressive, accelerator group of companies, while other firms are placed in a more conservative, non-accelerator category. If one is to have confidence that the postulated behavior patterns are meaningful and will persist, however, a link of some sort must be forged between behavior and character.

Suppose, for example, that one believes accelerator behavior to be more "aggressive" than residual funds behavior, and aggressive firms to be more "market" than "balance sheet" oriented. Suppose further that one expects such orientations to lead to more rapid rates of growth and lower liquidity positions for aggressive than for (relatively) conservative companies. Analyses of differences between within group growth-liquidity "profiles," then, may be used to evaluate the hypothesized relationship's credibility.

In addition to vague feelings of credibility, however, an ability to stratify observations into groups that display both profile and behavioral homogeneity also confers a number of very tangible analytic benefits; such as:

1. An ability to explain qualitative (as well as quantitative) differences in "between group" behavior,
2. An ability to forecast "within group" behavior with considerably improved precision, and
3. An ability to sort-out the effect on a population's aggregate behavior of differences in external conditions from differences in internal structure, or composition.

In a great many instances capabilities such as these may make the difference between success or failure, insight or frustration in applied econometric work.[4]

Profile Stratification

The need on occasion to stratify economic observations into behaviorally homogeneous groups is hardly new to the field of econometrics. Time series studies that segregate periods of war and depression from more normal times, or cross section analyses that stratify consumers by income, race, or sex, and firms by standard industrial classification, all are well known.

Also well known, however, is the fact that strata fine enough to produce reasonably homogeneous sets of observations generally retain either too few data elements or too little variation among behavioral variables in each cell to be of much experimental value; while classifications broad enough to be statistically viable fail to display either behavioral or profile homogeneity. The problem, of course, may be traced to the fact that

[4] G. H. Orcutt, et.al., Microanalysis of Socioeconomic Systems, Harper, New York, 1961.

conventional strata deal with 1,2,..., or at most a very few profile variates, one at a time. Unfortunately, each by itself is incapable of adequately representing the potentially complex interactions between profile dimensions required to differentiate basically different population types; yet together, the large number of data cells created makes extremely wasteful use of available statistical degrees of freedom. More effective means of objectively stratifying economic data clearly are needed.

One alternative, behavioral stratification, already has been suggested; and may be useful in certain applications. In a great many instances, however, behavioral alternatives may fail to be either sufficiently few in number or well defined to provide convenient, unambiguous, and exhaustive population subsets. Of more fundamental importance, behavioral stratification implies a willingness to make fairly strong prior assumptions about behavioral types, while profile differences are studied with greater scientific detachment. The main thrust of most econometric analyses, however, is exactly the reverse; beginning with relatively strong prior assumptions about relevant profile differences, economists tend to focus attention on explanations of observed behavior. Profile differences are <u>assumed</u> and behavioral differences are <u>studied</u>, rather than the other way around. Accordingly, strata based on profile rather than behavioral homogeneity are likely to hold the greater potential for econometric application.

Metric

An ability to group firms or other types of observations on the basis of profile similarity, of course, implies (or more accurately, <u>requires</u>)

an ability to measure similarity between observations in several dimensions. Returning to our earlier illustration, however, it quickly becomes apparent that such a requirement is non-trivial. How, for example, does one measure one firm's similarity to another, or either's similarity to a prototype of aggressive or conservative managerial temperament? Or more specifically, in the growth-liquidity profile space illustrated in Figure 1, how may one measure firm A's "nearness" to B, or either's "nearness" to C?

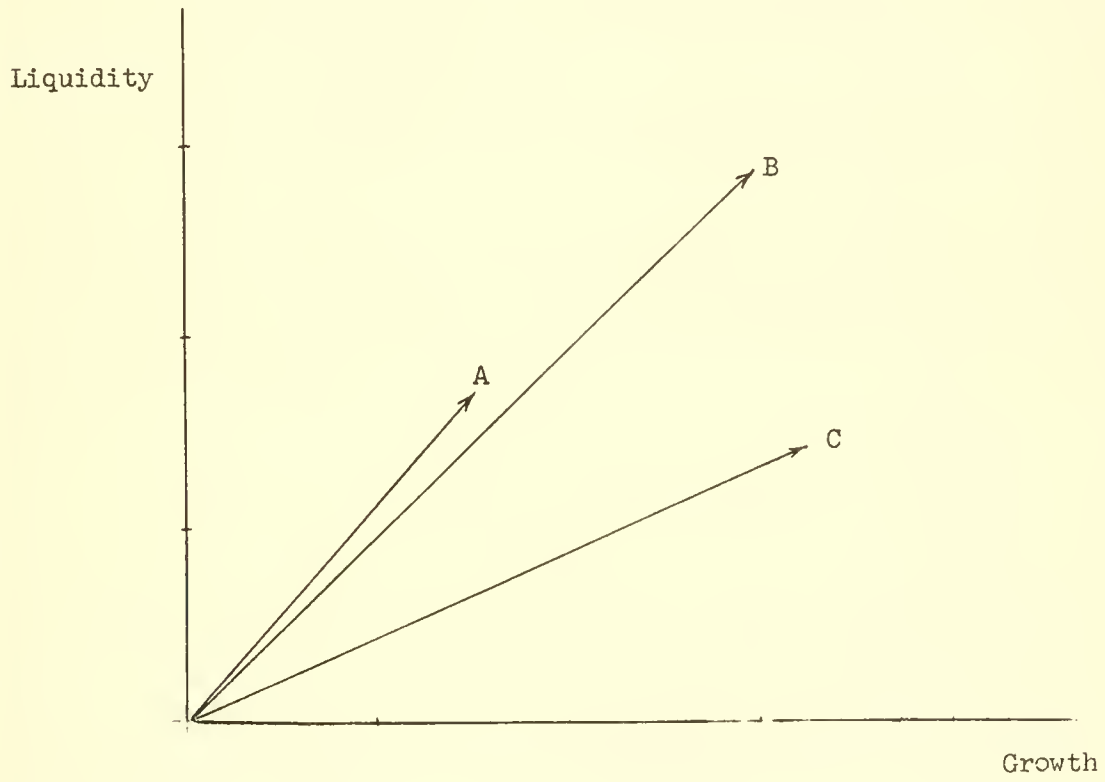

Figure 1

One obvious metric, clearly, can be provided by using a ruler to measure A's distance (as plotted) from B, C, or any other point in growth-liquidity space. The procedure carries a certain intuitive appeal, as well as the very practical capability of combining and reducing objective differences between pairs of observations in two dimensions to a single index. In addition, the concept of "distance in variables space" permits easy generalization to multiple dimensions. For example, defining:

$D^2_{i,j}$ as the squared distance between observations (or firms) i and j in profile space, and

$Z_{i,k}$ as observation i's value on the kth profile variable, k=1,2,...,p

a straightforward extension of the familiar Pythagorean theorem permits one to define distance-like measures of profile dissimilarity over as many characteristics as desired,

$$(3) \qquad D^2_{i,j} = \sum_{k=1}^{p} (Z_{i,k} - Z_{j,k})^2.$$

Similarity, however, is a subjective concept, and will correspond closely to mathematical measures such as (3) only if the variables through which profiles are measured are both comparably scaled and relatively independent of one another.

To illustrate, let us translate (3) into matrix notation. Defining:

$$\underline{d}_{i,j} = \begin{bmatrix} Z_{i,1} - Z_{j,1} \\ Z_{i,2} - Z_{j,2} \\ \vdots \\ Z_{i,p} - Z_{j,p} \end{bmatrix}$$

as a pxl vector of differences between observations i and j over k=1,2,...,p profile characteristics, and

$$\underline{\underline{I}} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

as a pxp identity matrix,

(3) can be written more compactly as

$$(4) \qquad D^2_{i,j} = \underline{d}^t_{i,j} \, \underline{\underline{I}} \, \underline{d}_{i,j}.$$

Examining (4) it becomes apparent that Pythagorean distance as usually defined is a very special case of a much broader class of distance-like measures, where multivariate profile differences $\underline{d}_{i,j}$ are projected through a transformation matrix $\underline{\underline{T}}$ into an index

$$(5) \qquad D^2_{i,j} = \underline{d}^t_{i,j} \, \underline{\underline{T}} \, \underline{d}_{i,j}$$

by which inter-profile differences may be reduced to a single criterion.

Generalized distance measures such as (5) offer tremendous flexibility to an analyst. $\underline{\underline{T}}$, clearly, can be chosen to impose any of a wide range of conditions on $D^2_{i,j}$. Variations in diagonal elements, for example, directly affect each variable's first order contribution to inter-firm differences; while off-diagonal elements may amplify, or modulate as desired, second order contributions to $D^2_{i,j}$ through the interaction of correlated profile variates.

An idea of the range of transformations available and of their assumptions, limitations, and implications for empirical work may be gained by examining three of the measures most frequently encountered in Psychometric applications. The first generally is attributed to Pearson, the second to Mahalanobis, and the third to Burt; although in each case the techniques have evolved sufficiently over time to give the original authors some difficulty recognizing the work attributed to themselves. Each satisfies the mathematical properties required of a metric space,[5] in that each provides both a set of points $\underline{d}_{i,j}$ in the space of real numbers, and a distance measure $D^2_{i,j}$ that is a non-negative, single-valued real function of $\underline{d}_{i,j}$ satisfying the following properties:

1. $D^2_{i,j} = 0$ if an only if all elements of $\underline{d}_{i,j}$ are null; i.e., only if $Z_{i,k} = Z_{j,k}$ for all $k=1,2,\ldots,p$;

2. $D^2_{i,j} = D^2_{j,i}$, the axiom of symmetry;[6] and

3. $D^2_{i,j} \leq D^2_{i,k} + D^2_{k,j}$, the triangle axiom.

In addition, all three commonly used metrics satisfy the requirement that distance (or dissimilarity) measures are non-decreasing functions of their arguments;[7] and two (those attributed to Mahalanobis and Burt) satisfy

[5] See for example, A. N. Kolmogorov and S. V. Fomin, Metric and Normed Spaces, Volume 1 of Elements of the Theory of Functions and Functional Analysis, Graylock Press, Rochester, N. Y., 1957; especially Chapter 2.

[6] Rao provides a fetching example of an instance in which the symmetry axiom may not be appropriate. Suppose, for example, that distance were to measure in some fashion the degree of "personal attractiveness" between persons i and j. Johnny (person i) may think Mary (person j) is wonderful, in which case $D^2_{i,j}$ is small, while Mary thinks Johnny is awful, so $D^2_{j,i} \neq D^2_{i,j}$ is large, and the axiom of symmetry is violated. This type of problem, although amusing, is not likely to undermine many econometric studies. See C. R. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," Journal of the Royal Statistical Society, Series B, X, 2, 1948.

[7] In the sense that adding additional variables to a profile will never increase the similarity, or decrease the distance between two observations.

the additional property that distance is a bounded function of its arguments -- i.e., that additional characteristics, or profile variables, contribute only independent information to measures of profile similarity.

Comparisons between Pearsonian and Mahalanobis distances are most easily developed in terms of uncorrelated rather than correlated variates. Following Harris' lead,[8] therefore, one may choose any of a wide range of alternative transformation matrices $\underline{\underline{T}}$, through which a given set of correlated profile variables $\underline{\underline{Z}}$ can be linearly decomposed into a corresponding set of uncorrelated variates, or factor scores, $\underline{\underline{F}}$. Any of an infinite number of transformations,

$$\underline{\underline{Z}} = \underline{\underline{F}}\ \underline{\underline{T}} \tag{6}$$

clearly are possible. For convenience, however, we will employ the method of principal components to obtain such a transformation.[9] Accordingly, let us define a set of profile variables $\underline{\underline{Z}}$, factor scores $\underline{\underline{F}}$, factor loadings $\underline{\underline{L}}$, and latent roots $\underline{\underline{\Lambda}}$, in such a fashion that:

$\underline{\underline{Z}}$ is an Nxp matrix of N observations on p profile variables, where each variable or column vector $\underline{Z}_j$ is scaled (by mean, standard deviation and sample size) to unit length; i.e., $\underline{Z}_j^t\ \underline{Z}_j = 1$. $\underline{\underline{Z}}^t\underline{\underline{Z}} = \underline{\underline{R}}$, accordingly, is a matrix of zero order correlation coefficients.

---

[8] C. W. Harris, "Characteristics of Two Measures of Profile Similarity," Psychometrika, 20, 4, December 1955.

[9] No attempt will be made here to reproduce proofs of the very convenient properties that lead us to employ Principal Components for this purpose. They will, simply, be summarized and used as necessary, below. For fuller discussions of the method see any of several standard references on principle components analysis, e.g., H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," Journal of Educational Psychology, 24, 1933; H. Harman, Modern Factor Analysis, University of Chicago Press, Chicago, 1960; M. G. Kendall, A Course in Multivariate Analysis, Hafner, New York, 1957; or T. W. Anderson, Introduction to Multivariate Analysis, Wiley, New York, 1958.

$\underline{\underline{F}}$ is defined as an Nxm matrix of m orthogonal factor scores on each of N observations, where $m \leq p$.

$\underline{\underline{L}}$ is an Nxm matrix of factor loadings, or latent vectors that project $\underline{\underline{Z}}$ into $\underline{\underline{F}}$. In addition, the matrix $\underline{\underline{L}}$ is internally orthogonal. Thus $\underline{\underline{L}}\underline{\underline{L}}^t = \underline{\underline{I}}_m$; and if $\underline{\underline{L}}$ is of full rank, $\underline{\underline{L}}^t\underline{\underline{L}} = \underline{\underline{I}}_p$ also holds.

$\underline{\underline{F}}^t\underline{\underline{F}} = \underline{\underline{\Lambda}}$ is an mxm diagonal matrix containing the latent roots that correspond to $\underline{\underline{L}}$. Each root $\lambda_i = \underline{F}_i^t \underline{F}_i$ generally is interpreted as the ith factor's variance.

Should $\underline{\underline{L}}$ be of full rank -- i.e., should as many artificial factors $\underline{\underline{F}}$ be derived as there are (non-singular) variables in $\underline{\underline{Z}}$ -- the method of principle components defines the exact transformation

$$\underline{\underline{Z}} = \underline{\underline{F}}\,\underline{\underline{L}}, \tag{7}$$

$$\underline{\underline{F}} = \underline{\underline{Z}}\,\underline{\underline{L}}^t. \tag{7a}$$

Should $\underline{\underline{L}}$ be less than full rank, however (i.e., should $m < p$ uncorrelated factors $\underline{\underline{F}}$ be defined) a residual or error matrix $\underline{\underline{U}}$ must be defined such that

$$\underline{\underline{Z}} = \underline{\underline{F}}\,\underline{\underline{L}} + \underline{\underline{U}}, \tag{8}$$

although

$$\underline{\underline{F}} = \underline{\underline{Z}}\,\underline{\underline{L}}^t \tag{8a}$$

continues to hold for those ($m < p$) factors that are defined.

Given a (complete) set of orthogonal factors $\underline{\underline{L}}$ capable of exactly reproducing the data in an observed set of profile variables $\underline{\underline{Z}}$, it also is capable of exactly reproducing the sample's matrix of zero order correlation coefficients,

$$\underline{\underline{R}} = \underline{\underline{Z}}^t\underline{\underline{Z}} = \underline{\underline{L}}^t\underline{\underline{F}}^t\,\underline{\underline{F}}\,\underline{\underline{L}} = \underline{\underline{L}}^t\,\underline{\underline{\Lambda}}\,\underline{\underline{L}}. \tag{9}$$

And finally, taking advantage of the orthogonality of $\underline{\underline{L}}$, the inverse matrix of zero order correlation coefficients, $\underline{\underline{R}}^{-1}$ (or indeed the inverse of any symmetric matrix) may be decomposed in highly simplified fashion into the quadratic form,

$$(10) \qquad \underline{\underline{R}}^{-1} = (\underline{\underline{L}})^{-1} \underline{\underline{\Lambda}}^{-1} (\underline{\underline{L}}^{t})^{-1} = \underline{\underline{L}}^{t} \underline{\underline{\Lambda}}^{-1} \underline{\underline{L}}.$$

These properties, although sufficient in a formal sense for our needs, hardly constitute a full description of the method of principle components. Like any factorial model, components analysis is designed not simply to reproduce data, but to interpret and to summarize certain characteristics of a sample in terms of a smaller number of relatively pervasive, underlying forces (or factors).[10] In particular, factorial techniques are designed to assist an analyst to sort out those portions of variance that appear to be common to a number of variates in $\underline{\underline{Z}}$ from those that are unique to particular variates, $\underline{Z}_i$. There is little doubt that a capability of this sort may be useful. Our need here, however, is simply for a convenient transformation [such as (7)] of correlated into uncorrelated profile variables; and for this purpose the method of principal components is both adequate and convenient.

Pearsonian Distance: Developed by Karl Pearson in 1926 as an inferential statistic to help physical anthropologists classify the racial origins of prehistoric skulls, the coefficient of Racial Likeness (CRL) has subsequently been developed (primarily by psychologists) into a descriptive

[10] Factorial methods, of course, are both the baby and the foundation of much of Psychometrics, as is regression analysis to econometrics. Controversy over the relative virtues of component vs. "purer" forms of factor analysis for identifying, summarizing, and interpreting the interdependent "structure" that underlies a set of intercorrelated variates $\underline{\underline{Z}}$ is widespread; and need not detain us here.

measure as well.[11] Like most such measures, Pearsonian distance can be developed and illustrated most easily through uncorrelated rather than correlated variables. Accordingly, let us transform a set of interdependent profile variates $\underline{\underline{Z}}$ into uncorrelated factor scores $\underline{\underline{F}}$ through a straightforward principal axes rotation,

$$\underline{\underline{Z}} = \underline{\underline{F}}\,\underline{\underline{L}}, \tag{11}$$

$$\underline{\underline{F}} = \underline{\underline{Z}}\,\underline{\underline{L}}^{t}, \tag{11a}$$

where $\underline{\underline{F}}$, $\underline{\underline{Z}}$, and $\underline{\underline{L}}$ all are as defined earlier. Similarly, interobservation profile differences can be defined either in terms of differences between rows i and j of $\underline{\underline{Z}}$ as before,

$$\underline{d}_{i,j} = \begin{bmatrix} Z_{i,1} - Z_{j,1} \\ Z_{i,2} - Z_{j,2} \\ \vdots \\ Z_{i,p} - Z_{j,p} \end{bmatrix}$$

or defining a new vector of (uncorrelated) interobservation differences $\underline{f}_{i,j}$, in terms of differences between corresponding rows of $\underline{\underline{F}}$, as

$$\underline{f}_{i,j} = \begin{bmatrix} F_{i,1} - F_{j,1} \\ F_{i,2} - F_{j,2} \\ \vdots \\ F_{i,m} - F_{j,m} \end{bmatrix},$$

[11] K. Pearson, "On the Coefficient of Racial Likeness," Biometrika, 18, 1926; L. J. Cronback and G. C. Gleser, "Assessing Similarity Between Profiles," Psychological Bulletin, 50, 1953; and C. W. Harris, op.cit.

where $\underline{d}_{i,j}$ and $\underline{f}_{i,j}$ are connected, through (11), by

$$\underline{d}^{t}_{i,j} = \underline{f}^{t}_{i,j} \underline{\underline{L}}, \tag{12}$$

and

$$\underline{f}^{t}_{i,j} = \underline{d}^{t}_{i,j} \underline{\underline{L}}^{t}. \tag{12a}$$

In general, we will assume that $\underline{\underline{Z}}$ is of full rank, and that a full set of m=p orthogonal factors are extracted and retained to characterize interprofile differences.

Employing the derived set of orthogonal variates, a simple Euclidean metric now may be employed to map vectors of multivariate differences $\underline{f}_{i,j}$ into unique measures of profile dissimilarity,

$$D^{2}_{i,j} = \underline{f}^{t}_{i,j} \underline{\underline{I}} \underline{f}_{i,j}. \tag{13}$$

From (12), however, it quickly becomes apparent that the transformation from correlated to uncorrelated profiles through a simple principal axes (or any other) rotation has no effect whatever on measured distance, for substituting (12) into (13), $D^{2}_{i,j}$ can be rewritten as

$$D^{2}_{i,j} = \underline{d}^{t}_{i,j} \underline{\underline{L}}^{t} \underline{\underline{I}} \underline{\underline{L}} \underline{d}_{i,j},$$

and recognizing the orthogonality of $\underline{\underline{L}}$, can be rewritten immediately as Pearson's familiar Coefficient of Racial Likeness,

$$D^{2}_{i,j} = \underline{d}^{t}_{i,j} \underline{\underline{I}} \underline{d}_{i,j}. \tag{14}$$

Being designed as an inferential statistic, Pearson's CRL presumes profile deviates to be normalized to zero mean, unit variance. As a descriptive measure, however, no such restriction is necessary; and Pearson's CRL

(14) is equivalent to the direct measurement of (squared) distances in the original space of one's profile variables; or to measuring distances between observations directly from Figure 1.[12]

Viewed in this fashion, the procedure's advantages are greatest, clearly, when a simple rotation rather than a projection of profile variables is desired. These are, of course, instances in which an analyst has a great deal of confidence in the space defined by his variables; i.e., when he is confident that the variables selected are reasonably uncorrelated, effectively span the space of meaningful profile characteristics, and are comparably scaled. These, of course, are ambitious requirements, and are likely to be satisfied only when a great deal of experience, strong prior convictions and a bit of luck are combined in a particular study.

Disadvantages of the CRL may be summarized under any of several headings; all, however, revolve about its treatment of intercorrelated profile variables. The problem, most simply stated, is that each variate's contribution to $D^2_{i,j}$ is not limited to its independent information. Accordingly, a set of 2,3,...,n variates, all of which measure essentially the same underlying pheonomenon, receives 2,3,...,n weight in distance measures; and tends to swamp by force of numbers whatever independent variation individual profile variates may possess. As a result, Pearson's metric is not bounded as a set of profile variables becomes infinite, even though all included variates are contained in a bounded space -- i.e., even though all are perfect linear combinations of a finite number of underlying components. The problem, of course, is compounded rather than ameliorated by dropping all except the first few "general

[12] If one's ruler is calibrated in squared rather than linear units.

factors" from F; for in such an instance only common dimensions of variance are retained, and these are the factors, or dimensions, most certain to explode as the number of (correlated) variates employed increases. Thus, similarities between skulls, all of which are humanoid, may swamp the rather subtle tribal differences sought by physical anthropologists,[13] just as similarities between firms, all of which are large, publicly held corporations, may swamp the relatively minor distinctions between aggressive and conservative managerial temperaments sought by econometricians. The larger and more highly correlated one's set of profile variables becomes, of course, the more vulnerable Pearson's "Coefficient of Racial Likeness" becomes as a measure of profile homogeneity.

Mahalanobis Distance: Mahalanobis' measure of interprofile distance attempts to counterbalance the overwhelming weight accorded correlated variates in the first few "general factors" of a principal axes rotation by deflating each derived factor by its own variance, or eigenvalue. Designed to deal with the problem of correlated variates, Mahalanobis' inverse transformation, like Pearson's principal axis transformation, can best be developed in terms of uncorrelated factor scores.[14] Instead of basing interprofile distances, as Pearson does, on differences between factor scores defined as

---

[13] See Cronback and Gleser, op.cit.

[14] See C. W. Harris, op.cit.; also, P. C. Mahalanobis, "On the Generalized Distance in Statistics," Proceedings of the National Institute of Sciences of India, 12, 1936; _____, "On Tests and Measures of Group Divergence, I," Journal of the Asiatic Society of Bengal, XXVI, 1930; and C. R. Rao, "Generalization of Markoff's Theorem and Tests of Linear Hypothesis," Sankhyā, 7, 9, 1945; _____, "The Utilization of Multiple Measurements in Problems of Biological Classification," Journal of the Royal Statistical Society, op.cit.; and _____, "Tests of Significance in Multivariate Analysis," Biometrika, 35, 1948.

$$\underline{\underline{F}} = \underline{\underline{Z}}\,\underline{\underline{L}}^{t}$$

where

$$\underline{\underline{F}}^{t}\underline{\underline{F}} = \underline{\underline{\Lambda}}\,,$$

Mahalanobis' distance is defined in terms of equally weighted factors,

$$\underline{\underline{F}} = \underline{\underline{Z}}\,\underline{\underline{L}}^{t}\,\underline{\underline{\Lambda}}^{-1/2}, \tag{15}$$

such that

$$\underline{\underline{F}}^{t}\underline{\underline{F}} = \underline{\underline{\Lambda}}^{-1/2}\,\underline{\underline{L}}\,\underline{\underline{Z}}^{t}\underline{\underline{Z}}\,\underline{\underline{L}}^{t}\,\underline{\underline{\Lambda}}^{-1/2} = \underline{\underline{\Lambda}}^{-1/2}\,\underline{\underline{L}}\,\underline{\underline{L}}^{t}\,\underline{\underline{\Lambda}}\,\underline{\underline{L}}\,\underline{\underline{L}}^{t}\,\underline{\underline{\Lambda}}^{-1/2} = \underline{\underline{I}}. \tag{16}$$

Defining interprofile differences between rows i and j of $\underline{\underline{Z}}$ and $\underline{\underline{F}}$ as $\underline{d}^{t}_{i,j}$ and $\underline{f}^{t}_{i,j}$, as before, and connecting them through (15), one obtains

$$\underline{f}^{t}_{i,j} = \underline{d}^{t}_{i,j}\,\underline{\underline{L}}^{t}\,\underline{\underline{\Lambda}}^{-1/2}, \tag{17}$$

and

$$\underline{d}^{t}_{i,j} = \underline{f}^{t}_{i,j}\,\underline{\underline{L}}\,\underline{\underline{\Lambda}}^{-1/2}. \tag{17a}$$

A straightforward Euclidean metric now can be used to map orthogonal, multivariate differences $\underline{f}_{i,j}$ into a unique and hopefully meaningful measure of interprofile dissimilarity,

$$D^{2}_{i,j} = \underline{f}^{t}_{i,j}\,\underline{\underline{I}}\,\underline{f}_{i,j}. \tag{18}$$

Employing (17), the metric's implications for the relative weights applied to one's original set of profile variables can easily be developed, by rewriting (18) as

$$D^{2}_{i,j} = \underline{d}^{t}_{i,j}\,\underline{\underline{L}}^{t}\,\underline{\underline{\Lambda}}^{-1}\,\underline{\underline{L}}\,\underline{d}_{i,j}. \tag{19}$$

Substituting (10) into (19), the inverse transformation

$$D^2_{i,j} = \underline{d}^t_{i,j} \underline{\underline{R}}^{-1} \underline{d}_{i,j}, \tag{20}$$

easily recognized as Mahalanobis' measure of interprofile distance, is obtained.

Several characteristics of Mahalanobis' measure should be noted. To begin, for normalized, uncorrelated profile variables, i.e., for

$$\underline{\underline{Z}}^t \underline{\underline{Z}} = \underline{\underline{R}} = \underline{\underline{R}}^{-1} = \underline{\underline{I}},$$

Mahalanobis and Pearsonian measures, (20) and (14), respectively, are identical. For correlated variates, of course,

$$\underline{\underline{R}}^{-1} \neq \underline{\underline{I}},$$

and the measures diverge -- generally quite dramatically.

While Pearson's CRL is directly sensitive to the scaling of one's variates [see (14)] Mahalanobis' index is quite insensitive to the scaling of $\underline{\underline{Z}}$. Defining a diagonal matrix of (positive) weights $\underline{\underline{W}}$ to be applied to a set of normalized (zero mean, unit variance) deviates, and factoring, one obtains

$$\underline{\underline{Z}}\,\underline{\underline{W}} = \underline{\underline{F}}\,\underline{\underline{L}}, \tag{21}$$

$$\underline{\underline{F}} = \underline{\underline{Z}}\,\underline{\underline{W}}\,\underline{\underline{L}}^t, \tag{21a}$$

and

$$\underline{\underline{L}}^t\,\underline{\underline{\Lambda}}\,\underline{\underline{L}} = \underline{\underline{W}}\,\underline{\underline{R}}\,\underline{\underline{W}}. \tag{22}$$

Following Mahalanobis, let us now deflate each vector of factor scores by $\underline{\underline{\Lambda}}^{-1/2}$. Vectors of correlated and uncorrelated interprofile differences accordingly, are related to one another through

$$(23) \qquad \underline{f}^{t}_{i,j} = \underline{d}^{t}_{i,j} \underline{\underline{W}} \underline{\underline{L}}^{t} \underline{\underline{\Lambda}}^{-1/2};$$

and Mahalanobis' D^2 measure,

$$(24) \qquad D^{2}_{i,j} = \underline{f}^{t}_{i,j} \underline{\underline{I}} \underline{f}_{i,j}$$

yields

$$(25) \qquad D^{2}_{i,j} = \underline{d}^{t}_{i,j} \underline{\underline{W}} \underline{\underline{L}}^{t} \underline{\underline{\Lambda}}^{-1} \underline{\underline{L}} \underline{\underline{W}} \underline{d}_{i,j} = \underline{d}^{t}_{i,j} \underline{\underline{R}}^{-1} \underline{d}_{i,j},$$

from which $\underline{\underline{W}}$ washes entirely.

Differences between Mahalanobis' and Pearson's measures of interprofile distance now can be stated quite succinctly. By assuming variables to be relatively independent and comparably scaled, Pearson produces a CRL that is highly sensitive both to the selection and the scaling of variables. Being dominated by the overwhelming weight accorded the first few general factors from a set of intercorrelated variates $\underline{\underline{Z}}$, Pearsonian distance ordinarily is not affected noticeably by an analyst's choice of factor structure, $\underline{\underline{F}}$.

By requiring each of the factors that underlies a set of profile variates to be comparably scaled (and, accordingly, to receive equal weight in measures of interprofile distance) Mahalanobis, on the other hand, produces a measure of interprofile distance that is quite insensitive to the selection of variates, and that is entirely insensitive to their scaling.[15] Weighting the least principal component as heavily as the greatest, of course, effectively neutralizes the impact on the first few general factors of a permissive approach

[15] As Rao notes, "One can be ... safe ... using D^2 with some superfluous [variables]. If any [variable] produces an appreciable change in D^2, it is taken to be additional value in discrimination." See C. R. Rao, "The Utilization of Multiple Measurements in Problems of Biological Classification," Journal of the Royal Statistical Society, B, X, 1948.

to the selection of variates from highly intercorrelated alternative measures. Unfortunately, such a procedure also blows-up any random "error" that later factors may contain; threatening to swamp meaningful dimensions of information with pure statistical noise.

The successful use of Mahalanobis' measure, accordingly, is likely to depend not on one's selection and scaling of variates, but instead on his choice of factor structure. In general, such a structure should not be full, but should contain fewer than p carefully derived, equally weighted, orthogonal factors, $\underline{F}$; and in this sense will differ from that originally proposed by Mahalanobis (20). For finite m, of course, such a measure would be bounded even over an infinite set of profile variates. The (modified) measure's robustness, however, does not eliminate the need for a certain amount of artistry, as well as craftsmanship, in the development of distance measures -- for both generally are required to decompose an observed set of correlated profiles $\underline{Z}$ into a meaningful set of uncorrelated factors, $\underline{F}$.

Q Correlation: Multivariate measures of profile similarity, of course, need not be based exclusively on Euclidean distance measures. The oldest, and perhaps still the most commonly used non-Euclidean measure is not strictly speaking a measure of distance at all; but instead is a measure of inter-person, or interprofile correlation.[16]

Correlation, of course, ordinarily is thought of as a measure of the angular separation between variables; specifically, as the cosine of an angle that separates a pair of variables in observation space, or as the cosine of

[16] For references to the classical literature on Q-Correlation see, C. Burt, "Correlations Between Persons," British Journal of Psychology, 28, 1937. For empirical applications, see W. Stephenson, The Study of Behavior, University of Chicago Press, Chicago, 1953; and for a critical discussion, see J. L. Cronback and G. C. Gleser, op.cit.

the angle between column vectors of $\underline{Z}$. Should $\underline{Z}$ be normalized (to zero mean, unit variance) by columns, the inner product

$$\underline{Z}^t\underline{Z} = \underline{R} \tag{26}$$

defines a pxp matrix of zero order correlation coefficients between variables.

Q-correlation is similarly defined, except that angular separation is measured between observations in variables space, or between rows of $\underline{Z}$ rather than variables, or columns of $\underline{Z}$. Accordingly, let us define a transformation $\underline{\xi}$ of $\underline{Z}$ such that $\underline{\xi}$ is normalized to zero mean unit variance by rows. A (perhaps very large) NxN matrix $\underline{Q}$ of zero order correlation coefficients between observations now can be defined as the outer product,

$$\underline{\xi}\,\underline{\xi}^t = \underline{Q}. \tag{27}$$

Developed largely by and for psychologists, Q correlation's geometric properties, including its relation to Euclidean distance, may be illustrated graphically through an example such as Figure 2; where vectors A and B represent scores by persons on batteries of arithmetic and verbal aptitude tests plotted in the space defined by the tests (or variables). Pearsonian distance between A and B, clearly, can be measured directly from the Figure as the length of a line (not shown) connecting the points. Q correlation measures angular distance between persons through the zero order correlation coefficient, $q = \cos\ \theta$.

Normalizing observations to unit length, of course, is equivalent to projecting A and B into A' and B' on a unit circle from the origin. q accordingly, can be measured by the perpendicular (canonical) projection of

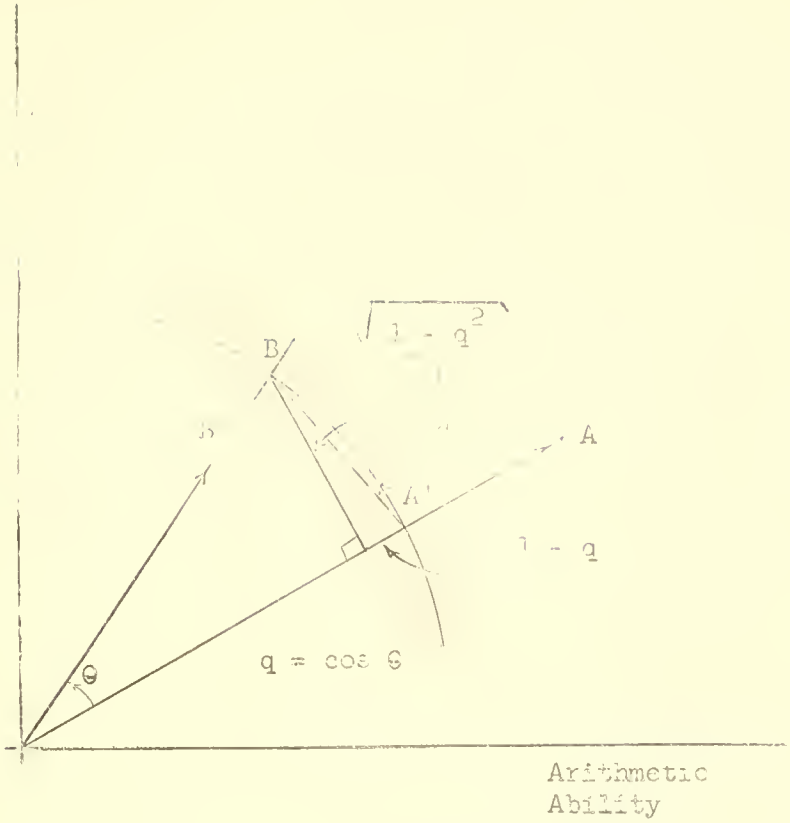

Figure 2

B' onto A (or, A' onto B); and can easily be seen from the figure to vary inversely with the squared distance between normalized vectors A' and B',[17] as

$$d^2 = 2(1 - q). \qquad (28)$$

[17] With Stephenson, and Cronback and Gleser as notable exceptions, much of the psychometric literature surrounding Q correlation tends to play down the significance of the standardization (within observations, or rows of $\underline{Z}$). Burt in particular argues that similarities between persons and variables, analyzed through Q and R correlation, respectively, "should lead to consistent, and in the end, to identical conclusions." This proposition is anything but self-evident. See for example, L. Cronback and G. Gleser, op.cit.; and W. Stephenson, "Correlation Persons Instead of Tests," Character and Personality, 4, 1935; _____, "The Significance of Q-technique for the Study of Personality," in M. Reymert, editor, Feelings and Emotions, McGraw-Hill, 1950; and _____, "Some Observations on Q Technique," Psychological Bulletin, 49, 1952.

By standardizing each observation to unit length, Q correlation loses (indeed, throws out) the type of information generally summarized in the first few common factors of conventional component or factor analyses -- general ability (or IQ) in Psychometric studies, and technological or scale induced commonalities between firms in econometric data. For some purposes such a standardization may be eminently sensible, for other purposes, nonsensical. In any case, it seldom is trivial. It also appears clear from (28) that, given the implicit standardization $\underline{\underline{S}}$ of $\underline{Z}$, Q correlation offers no particular advantage over more conventional measures of interprofile distance.

Summary

No choice of metric in a multivariate profile space ever is likely to be unassailable. Each has its own set of properties, its own sensitivity to information requirements, and as such its own potential and limitations in any specific empirical context. Should standarized observations be desired for a particular study, Q correlation would appear to offer an appropriate metric. Should standardization across variables, or linear combinations of variables be preferred, Pearsonian or Mahalanobis distance-like measures become attractive.

Should one be uncomfortable over his selection of variables -- fearing, perhaps that superfluous measures may unduly affect the weight attached to underlying profile attributes -- and prefer to hang his hat on a specific factorization of $\underline{Z}$, Mahalanobis' D^2 provides an appropriate metric. Should one, on the other hand, have confidence in both his selection and scaling of variables, a simple Pearsonian metric that preserves these properties usually will be preferred. The author has experimented with all three

metrics over the past few years and is satisfied that each, in fact, possesses the properties (i.e., the sensitivities) advertised, and that each accordingly deserves its special place among the (infinite) array of alternative available measures of multivariate profile similarity.

It also is apparent that, in most instances, the stratification of observations $\underset{=}{Z}$ into groups or subsets $\underset{=}{Z}^{(s)}$ will be highly sensitive to one's choice of metric;[18] and that, when all is said and done, one's best defense (or criticism) of any particular metric is likely to be his evaluation of the types and the usefulness of the groups that it produces. Are the groups interpretable? Do they make sense intuitively? And, of course, do the different strata of persons, households, firms, industries, years, etc. produced, display noticeably different patterns of observed behavior? Ongoing empirical studies tend to answer each of these questions in the affirmative.[19]

---

[18] K. J. Jones, "Problems of Grouping Individuals, and the Method of Modality," Harvard University Graduate School of Education, and Brandeis University, unpublished M.S., 1965.

[19] See for example, D. E. Farrar, R. R. Glauber and J. R. Meyer, "Patterns of Corporate Financial Behavior: An Empirical Analysis," paper delivered to a joint session of the American Finance Association and Econometric Society, Christmas Meetings, New York, 1965.

References

Anderson, T. W., Introduction to Multivariate Analysis, Wiley, New York, 1958.

Burt, C. L., "Correlation Between Persons," British Journal of Psychology, 28, 1937.

Cronback, L. J., and Gleser, G. C., "Assessing Similarity Between Profiles," Psychological Bulletin, 50, 6, 1953.

Eisner, R., "A Permanent Income Theory for Investment," forthcoming in the American Economic Review.

_____, and R. H. Strotz, "Determinants of Business Investment," Impacts of Monetary Policy, Commission on Money and Credit, Prentice-Hall, 1963.

Farrar, D. E., R. R. Glauber, and J. R. Meyer, "Patterns of Corporate Financial Behavior: An Empirical Analysis," paper delivered to a joint session of the American Finance Association and Econometric Society, Christmas Meetings, New York, 1965.

Graybill, F., An Introduction to Linear Statistical Models, Vol. 1, McGraw-Hill, New York, 1961.

Harman, H., Modern Factor Analysis, University of Chicago Press, Chicago, 1960.

Harris, C. W., "Characteristics of Two Measures of Profile Similarity," Psychometrika, 20, 1955.

Hotelling, H., "Analysis of a Complex of Statistical Variables into Principal Components," Journal of Educational Psychology, 24, 1933.

Jones, K. J., "Problems of Grouping Individuals, and the Method of Modality," Harvard University Graduate School of Education, and Brandeis University, unpublished M.S., 1965.

Kendall, M. G., A Course in Multivariate Analysis, Hafner, New York, 1957.

Kolmogorov, A. N., and S. V. Fomin, Metric and Normed Spaces, Volume 1 of Elements of the Theory of Functions and Functional Analysis, Graylock Press, Rochester, N. Y., 1957.

Mahalanobis, P. C., "On Tests and Measures of Group Divergence," Journal of the Asiatic Society of Bengal, 26, 1930.

_____, "On the Generalized Distance in Statistics," Proceedings of the National Institute of Science in India, 12, 1936.

Meyer, J. R., "On the Measurement of Business Motivation," Review of Economics and Statistics, 49, 3, 1967.

Orcutt, G. H., et.al., Microanalysis of Socioeconomic Systems, Harper, New York, 1961.

Pearson, K., "On the Coefficient of Racial Likeness," Biometrika, 18, 1926.

Rao, C. R., "Generalization of Markoff's Theorem and Tests of Linear Hypothesis," Sankhya, 7, 9, 1945.

_____, "Tests of Significance in Multivariate Analysis," Biometrika, 35, 1948.

_____, "The Utilization of Multiple Measurements in Problems of Biological Classification," Journal of the Royal Statistical Society, Series B, 10, 1948.

Stephenson, W., "Correlation Between Persons Instead of Tests," Character and Personality, 4, 1935.

_____, "Some Observations on Q Technique," Psychological Bulletin, 49, 1952.

_____, "The Significance of Q Technique for the Study of Personality," in M. Reymert, Editor, Feelings and Emotions, McGraw-Hill, 1950.

_____, The Study of Behavior, University of Chicago Press, 1953.

Theil, H., Economic Forecasts and Policy, 2nd Edition, North Holland, Amsterdam, 1961.

_____, Linear Aggregation of Economic Relations, North Holland, Amsterdam, 1954.

BASEMENT

Date Due

AUG 15 '72		
[illegible] 73		
APR 26 '83		
[illegible]		

Lib-26-67

MIT LIBRARIES
3 9080 003 901 474
244-67

MIT LIBRARIES
3 9080 003 901 409
245-67

MIT LIBRARIES
3 9080 003 901 433
246-67
HD28
.M414
Nos.2[illegible]-
Nos.253-[illegible]

MIT LIBRARIES
3 9080 003 901 326
247-67

MIT LIBRARIES
3 9080 003 901 391
248-67

MIT LIBRARIES
3 9080 003 870 471
249-67

MIT LIBRARIES
3 9080 003 870 356
250-67

MIT LIBRARIES
3 9080 003 901 383
251-67

MIT LIBRARIES
3 9080 003 901 367
252-67

MIT LIBRARIES
3 9080 003 870 398
253-67